NPS-OR-92-013

NAVAL POSTGRADUATE SCHOOL

Monterey, California

INTERACTIVE ANALYSIS OF GAPPY BIVARIATE TIME SERIES USING AGSS

Peter A. W. Lewis and Bonnie Ray

June 1992

Prepared for:
National Research Council
2101 Constitution Ave.,
Washington DC 20418

NAVAL POSTGRADUATE SCHOOL,
MONTEREY, CALIFORNIA

Rear Admiral R. W. West, Jr.
Superintendent

Harrison Shull
Provost

This report was funded by the National Research Council 2101 Constitution Ave., Washington, DC, 20418.

This report was prepared by:

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NPSOR-92-013	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Postgraduate School 6b. OFFICE SYMBOL (*If applicable*) OR	7a. NAME OF MONITORING ORGANIZATION Naval Postgraduate School
6c. ADDRESS (*City, State, and ZIP Code*) Monterey, CA 93943	7b. ADDRESS (*City, State, and ZIP Code*) Monterey, CA 93943-5006
8a. NAME OF FUNDING/SPONSORING ORGANIZATION National Research Council 8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER O&MN Direct Funding
8c. ADDRESS (*City, State, and ZIP Code*) National Research Council 2101 Constitution Ave., Washington DC 20418	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO / PROJECT NO / TASK NO / WORK UNIT ACCESSION NO

11. TITLE (*Include Security Classification*)
Interactive Analysis of Gappy Bivariate Time Series using AGSS

12. PERSONAL AUTHOR(S)
Peter A. W. Lewis and Bonnie Ray

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, month day*)	15. PAGE COUNT
Final Report	FROM 9/91 TO 5/92	June, 1992	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	Time series; interpolation; bivariate

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Bivariate time series which display nonstationary behavior, such as cycles or long-term trends, are common in fields such as oceanography and meteorology. These are usually very large-scale data sets and often may contain long gaps of missing values in one or both series, with the gaps perhaps occurring at different time periods in the two series. We present a simplified but effective method of interactively examining and filling in the missing values in such series using extensions of the methods available in AGSS, an APL2-based statistical software package. Our method allows for possible detrending and removal of seasonal components before automatically estimating arbitrary patterns of missing values for each series. Interactive bivariate spectral analysis can then be performed on the detrended and deseasonalized interpolated data if desired. We illustrate our results using a bivariate time series of ocean current velocities measured off the California coast.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT [x] UNCLASSIFIED/UNLIMITED [ ] SAME AS RPT. [ ] DTIC USERS	21. ABSTRACT SECURITY CLASSICIATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL P. Lewis	22b. TELEPHONE (*Include Area Code*) (408) 646-2283 22c. OFFICE SYMBOL OR/Lw

Interactive Analysis of Gappy Bivariate Time Series Using AGSS

Peter A. W. Lewis and Bonnie K. Ray
Dept. of Operations Research, Naval Postgraduate School
Monterey, CA 93943

Abstract

Bivariate time series which display nonstationary behavior, such as cycles or long term trends, are common in fields such as oceanography and meteorology. These are usually very large scale data sets and often may contain long gaps of missing values in one or both series, with the gaps perhaps occurring at different time periods in the two series. We present a simplified but effective method of interactively examining and filling in the missing values in such series using extensions of the methods available in AGSS, an APL2-based statistical software package. Our method allows for possible detrending and removal of seasonal components before automatically estimating arbitrary patterns of missing values for each series. Interactive bivariate spectral analysis can then be performed on the detrended and deseasonalized interpolated data if desired. We illustrate our results using a bivariate time series of ocean current velocities measured off the California coast.

1 Introduction

Gaps of missing values of various sizes are a common problem in many data sets. In oceanographic data, for example, a single large gap may arise in the gathering of tidal data when an instrument stops working and the malfunction is not detected for several days. Many small gaps are more characteristic of data gathered from satellites. The missing value problem is complicated for bivariate series in that the gaps may not fall at the same time periods in both series. Ad hoc univariate methods, such as basing "suitable" replacement values on the range of values assumed by neighboring points or points of the same periodicity, fail to account for possible cross correlation in the data. In order to successfully analyze the spectrum of gappy data sets, or use the data for other purposes, the missing values need to be estimated in a way that is characteristic of the rest of the bivariate data set.

The problem of missing values in time series has been studied by several authors in recent years, primarily in a state space framework. Jones (1980) used a Kalman filter recursion to calculate the exact likelihood of a univariate stationary autoregressive moving average (ARMA) process with missing values, while Harvey and Pierce (1984) and Kohn and Ansley (1986) extended the Kalman filtering method to nonstationary autoregressive integrated moving average (ARIMA) processes. Ansley and Kohn (1985) gave a method of recursively calculating the likelihood for a multivariate state space model with incompletely specified initial conditions which can be used to interpolate an arbitrary pattern of missing values in multivariate time series. Both the computation and derivation are much simpler in the univariate case than in the multivariate case. More recently, Ljung (1989) derived an exact expression for the estimates of missing values in a univariate ARIMA process in a form that is useful for examining the estimates and their mean squared errors. For an arbitrary pattern of missing values, however, the computations are not very efficient. None of the above methods is thus easy to implement in practice for bivariate series which are possibly nonstationary and have arbitrary patterns of missing values in both series.

In this paper, we present an algorithm for semi-automatically filling in gaps in bivariate time series, allowing for trends, cycles, and cross correlation in the data. The interactive implementation of the algorithm allows for visual examination of the data at each step, giving the practitioner the opportunity to view the original data with missing values, the "patched" data in which the missing values have been filled in using linear interpolation, and the estimated autospectrum of the crudely interpolated data. After this examination, one may choose to remove a trend or cycles from the data. The remaining series is automatically modeled as an autoregressive process and the estimated model is used for interpolation. The method allows

for joint interpolation of two correlated series, incorporating an estimate of the autocorrelation for each series, and the cross correlation between the two series, into the interpolated values. At the end of the interpolation phase, the user has the choice of examining the coherence function of the interpolated series, as well as producing more detailed plots of particular segments of the series. The following section presents the interpolation algorithm in detail, while Section 3 gives an application of the algorithm to a bivariate series of ocean current velocity meter readings measured off the California coast.

2 Algorithm

The following algorithm has been coded in APL2 using the IBM APL2 AGSS program as a computing platform. Thus functions such as regression, the Fast Fourier Transform used for computing the periodogram of the series, and random number generation from AGSS are used, as well as some of the AGSS graphics screens. The algorithm is available in a AGSS library from the authors for mainframe or microcomputer data. The algorithm is as follows:

1. The user is asked to enter the name of the original series containing gaps and the series is plotted. Denote this series by $x(t)$, $t = 1, 2, \ldots, n$. If there are two series containing gaps, and the two series are cross correlated in some way, the user is asked to enter the name of the second series and the second series is plotted as well. Denote this series by $y(t)$, $t = 1, 2, \ldots, n$. The two series must have the same length, but the location and length of gaps in the series may differ.

2. The user is asked to enter the number used on the data record to indicate a missing value.

3. The program then computes the locations of the gaps and fills in the missing values for each series by linearly interpolating between the two points on either side of the gap. Denote the linearly interpolated series as $x_1(t)$ and $y_1(t)$ respectively. The resulting series are then plotted and can be visually examined by the user to decide whether removal of a linear trend is necessary.

   Note: Steps 4–8 are applied to both the $x_1(t)$ and $y_1(t)$ series in exactly the same manner. Only the results for the $x_1(t)$ series are given below.

4. If so desired, the program removes a linear trend from each series. The trend is estimated using least squares applied to the initial "patched" series. The resulting series is

   $$x_2(t) = x_1(t) - \hat{a} - \hat{b}t,$$

   where $\hat{a}$ is the estimated constant and $\hat{b}$ is the estimated slope.

5. The sample periodogram is automatically estimated and plotted for the interpolated and detrended data, $x_2(t)$. (see, for example, Priestly (1981) for a definition of the periodogram and its interpretation).

6. The program calculates the probability of obtaining the computed values for the 20 largest values of the normalized periodogram under the assumption that $x_2(t)$ is a Gaussian white noise process. A small probability indicates that a cycle may be present in the data. The probabilities are computed using the large sample test statistic for $I_j, j = 1, 2, [n/2]$, where I_j denotes the j^{th} largest ordinate of the periodogram. (Priestly 1981, p.407).

7. Using the information obtained in the previous step, as well as any intuitive or physical knowledge of cyclic behavior in the series, the user specifies cycles to be estimated and removed from the interpolated and detrended data, if desired. The cycles are assumed to be of the form

   $$s_t = \sum_{j=1}^{J} \{\gamma_j cos(\omega_j t) + \beta_j \sin(\omega_j t)\},$$

   where $\omega_j = 2\pi f_j$ are the frequencies that you would like to remove and J is the number of cycles. The γ_j and β_j are estimated using least squares. The resulting series is $x_3(t) = x_2(t) - \hat{s}_t$.

8. A first order autoregressive (AR) model is fitted to the detrended and deseasonalized data $x_3(t)$. An AR(1) model has the form

   $$x_3(t) = \varphi x_3(t-1) + a_x(t), \quad t = 2, 3, \ldots, n.$$

   We assume that $a_x(t) \sim iid\, N(0, \sigma^2_{a_x})$. The parameter φ is estimated using least squares. The residual series $\hat{a}_x(t)$ is

   $$\hat{a}_x(t) = x_3(t) - \hat{\varphi} x_3(t-1), \quad t = 2, 3, \ldots, n.$$

9. The variance of $\{\hat{a}_x(t)\}$ is calculated and a white noise series of length n having the distribution $N(0, \hat{\sigma}^2_{a_x})$ is generated. Denote this series by $a'_x(t)$.

10. Let l denote the length of a particular gap in the series and let $x_3(t)$ and $x_3(t+l+1)$ denote the points on either side of the gap. The program forecasts and backcasts from each end of the gap using the following recursive equations for $j = 1, 2, \ldots, l$.

$$\begin{aligned} \hat{x}_3(t+j) &= \hat{\phi}\hat{x}_3(t+j-1) + a'_x(t+j) \\ \hat{x}_3(t+l+1-j) &= \hat{\phi}x_3(t+l+2-j) \\ &\quad + a'_x(t+l+1-j). \end{aligned}$$

Then the interpolated value is

$$\tilde{x}_3(t+j) = w_{1,j}\hat{x}_3(t+j) + w_{2,j}\hat{x}_3(t+l+1-j),$$

where $w_{1,j} = 1 - (j/l+1)$ and $w_{2,j} = 1 - w_{1,j}$.

11. If interpolating values for two correlated series, the standard deviation of the residual series $\{\hat{a}_x(t)\}$ and $\{\hat{a}_y(t)\}$ found in Step 8 is calculated and the sample cross correlation at lag 0 between $\{\hat{a}_x(t)\}$ and $\{\hat{a}_y(t)\}$ is computed. Denote this by c. A white noise series of length n having the distribution $N(0, \hat{\sigma}^2_{a,y})$ is generated. Denote this series by $a'_y(t)$. A second white noise series of length n is generated using the following relation:

$$a''_y(t) = c(\hat{\sigma}_{a,y}/\hat{\sigma}_{a,x})a'_x(t) + \sqrt{1-c^2}\,a'_y(t).$$

12. The values for the $y_3(t)$ series are interpolated as in Step 9, with $a'_x(t+j)$ replaced by $a''_y(t+j)$

13. The estimated trend and cycle are added back to the interpolated series and the series containing the final estimates of the missing values is plotted.

14. The user may choose to plot the coherence function for the detrended and deseasonalized interpolated series if desired.

15. The user may also choose to plot more detailed segments of the final interpolated series if desired.

Using a weighted average of backward and forward forecasts made from each end of a gap using an estimated univariate ARMA model, as was used in Step 10, was discussed by Abraham (1981). He calculates the weights to minimize the mean-squared error of the interpolated value, thus the weights depend in a complicated way on both the estimated model parameters and the length of gap. Our method, although simple, is intuitively appealing in that it gives less weight to interpolated values at long lead times and is easy to implement when the lengths of the gaps are different. Note that in Step 10, we include a simulated noise term in the usual expressions for the backward and forward forecasts of an AR(1) model. This is to eliminate unrealistic "smoothness" in the interpolated values, which will occur if the gap is very long. Realistic noise in the series is important if the interpolated series will be used to estimate the spectral density of the series. Similarly, in Step 11 we incorporate an estimate of the contemporaneous correlation between the two series into the estimates of the interpolated values of the second series by including a noise term taken from a simulated bivariate Normal pair of series with correlation c. The method for generating a bivariate Normal pair with specified correlation is taken from Lewis and Orav (1989, p.301). A discussion of contemporaneous bivariate time series models made be found in Camacho, Hipel, and McLeod (1987).

3 An Example

We apply the above algorithm to a vector pair of ocean current velocities collected off Point Sur, California over the period 0000 hours, Sept. 19, 1990 to 2300 hours, Oct. 30, 1990, a total period of 1008 hours. Current velocities are just one set of variables which are collected regularly by oceanographers at the Naval Postgraduate School in order to provide information related to the long term variability in sea surface temperatures off the California coast. The velocities were measured using a paddlewheel and electronic counter assembly located at the top of the recording unit placed at a depth of 350 meters. Velocity, in units of cm/s, was determined from the number of revolutions made by the paddlewheel during each sampling interval, to an accuracy of ± 1.0 cm/s. The data was initially recorded at 30 minute intervals. After initial visual inspection for outliers or periods suspect of instrument failures, and manual editing if necessary, the data was filtered using a Cosine-Lanczos filter with a centered 25 point data window and interpolated to specified 60 minute intervals. At this point, there remained a period of 63 consecutive hours of missing data, in which the data gathering instruments were not working. There were also a few scattered individual missing values.

Figures 1 and 2 show plots of the E-W (or u) component of the current velocity, and the N-S (or v) component of the current, with missing values coded as 0. Figures 3 and 4 depict the same series with missing values "patched" using linear interpolation. There appear to be regular cycles in the data, as expected for current data, as well as the presence of a long term trend. We fit a linear trend to both the u and v cur-

rent components, with estimated constants of 3.91 and 1.60, respectively, and estimated slope coefficients of -0.0018 and -0.0003. Standard t-tests on the significance of the regression coefficients are not appropriate because the residuals are not assumed to be uncorrelated. Figures 5 and 6 show the periodogram for each detrended series. Diurnal and semi-diurnal cycles are clearly indicated for the v-component of current velocity, with only the diurnal cycle clearly seen in the u-component. In addition, there appears to be some long range dependence in the data, as evidenced by the large values of the periodogram at small frequencies, which remain even after the removal of a long term trend. See Lewis and Ray (1992) for a discussion of long range dependence in sea surface data. Based on the approximate p-values of the test statistic for each of the 20 largest periodogram ordinates and knowledge of the tidal cycles, we remove cycles at frequencies (in cycles per 1008 hours) 81, 82, and 84 in the u-component and frequencies 42, 81, 82, and 84 in the v-component. Table 1 gives the values of the normalized periodogram values at these ordinates and the resulting p-values for the test statistic. After this step, the missing values are automatically estimated for the detrended and deseasonalized data. Figures 7 and 8 show the two series with final estimates of the missing values, after the trend and cycles have been added back to the series. The estimated values appear to follow the pattern of the data quite nicely.

We study the correlation between the two series in more detail by looking at the coherence function for the two interpolated, detrended and deseasonalized series. The coherence is initially computed assuming a cosine arch window of length 7. The user has the option of changing the window at any point in the coherence analysis. Figure 9 shows the cross-phase, cross-amplitude, and cross-coherence functions of the two series. The cross-amplitude spectrum shows dependence between the series at fairly low frequencies, but the (normalized) coherence measure shows that the dependence extends to high frequencies as well. No systematic effect can be seen in the phase spectrum.

Additionally, the enlarged segment in Figure 10 depicts the two series with values plotted as vertical lines drawn from the x-axis. The segment partially includes the interpolated values shown in Figures 7 and 8. No apparent difference between the known and the interpolated values can be seen.

4 Summary

We have presented a simple algorithm which permits interpolation of arbitrary patterns of missing values in both univariate and bivariate time series, allowing for the possibility of non-stationarity. The implementation is interactive and has graphical capabilities available at each step. It is also much easier to implement in practice than the state-space approach of Ansley and Kohn(1985), which requires modified versions of the Kalman filter and the fixed point smoothing algorithm. The estimated contemporaneous correlation between the two series is used in the interpolation algorithm in order to estimate the missing values in a manner that is consistent with the rest of the data. Although we have assumed that each series follows a simple AR(1) model, the algorithm could easily be extended to model each series as an ARMA(p, q) model, with the orders of p and q chosen by the user after examination of the sample autocorrelation and partial autocorrelation functions of the detrended and deseasonalized "patched" data. Functions necessary to compute the sample correlation functions and estimate the parameters of an ARMA(p, q) model are already present in the IBM APL2 AGSS package. (The AGSS application disdussed here is available from the authors; the AGSS package is available from IBM.)

References

Abraham, B. (1981), "Missing observations in time series", *Commun. Statist.–Theor. Meth.*, A10(16), 1643–1653.

Ansley, C. F. and Kohn, R. (1985), "Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions," *Annals of Statistics*, 13, 1286–1316.

Camacho, F., McLeod, A. I., and Hipel, K. W. (1987) "Contemporaneous bivariate time series," *Biometrika*, 74(1), 103-113.

Harvey, A. C., and Pierse, R. G. (1984), "Estimating missing observations in economic time series," *Journal of the American Statistical Association*, 79, 125–131.

Jones, R. H. (1980), "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, 22, 389–395.

Kohn, R. and Ansley, C. F. (1986), "Estimation, prediction, and interpolation for ARIMA models with missing data," *Journal of the American Statistical Association*, 81(395), 351–361.

Lewis, P. A. W. and Orav, E. J. (1989) *Simulation Methodology for Statisticians, operations Analysts, and Engineers*, Pacific Grove: Wadsworth & Brooks/Cole.

Lewis, P. A. W. and Ray, B. K. (1992), "Modeling nonlinear time series with long range dependence," Technical report, Naval Postgraduate School.

Ljung, Greta M. (1989), "A note on estimation of missing values in time series," *Commun. Statist. – Simula.*, 18(2), 459–465.

Priestley, M. B. (1981) *Spectral Analysis and Time Series*, London: Academic Press.

Tables and Figures

Table 1: Normalized Periodogram Values for Ocean Current Velocities

U-component of current velocity		
Frequency	Norm. Periodogram Value	p-value
1	106.14	0.00
81	88.52	0.00
2	35.67	0.00
82	22.42	0.00
84	20.85	0.00
V-component of current velocity		
Frequency	Norm. Periodogram Value	p-value
81	195.51	0.00
1	69.88	0.00
84	34.81	0.00
2	20.90	0.00
82	14.40	0.00
42	10.09	0.11

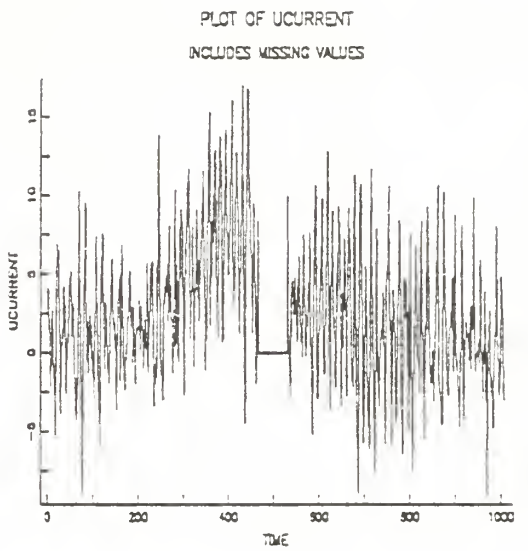

Figure 1: U-component of Current Velocity (missing values coded as 0's)

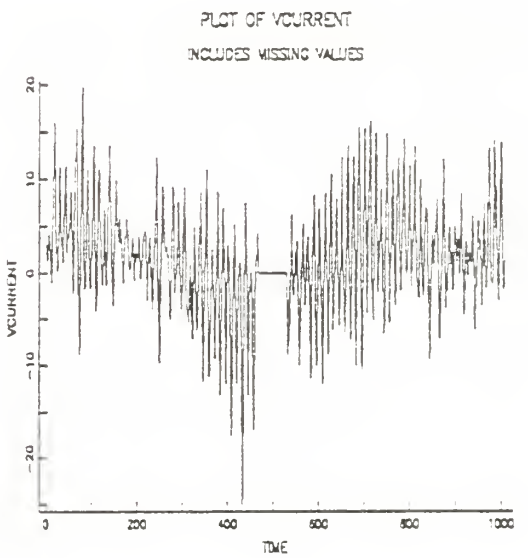

Figure 2: V-component of Current Velocity (missing values coded as 0's)

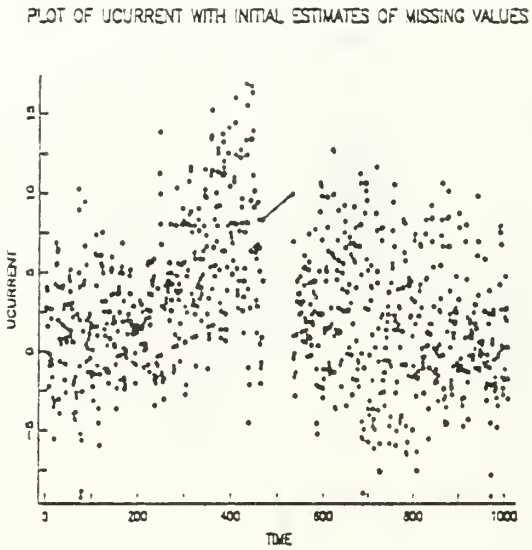

Figure 3: U-component of Current Velocity (missing values linearly interpolated)

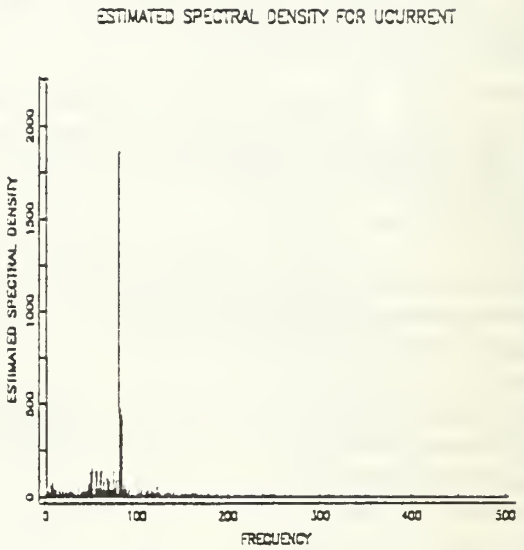

Figure 5: Periodogram of U-component of Current Velocity after linear detrending

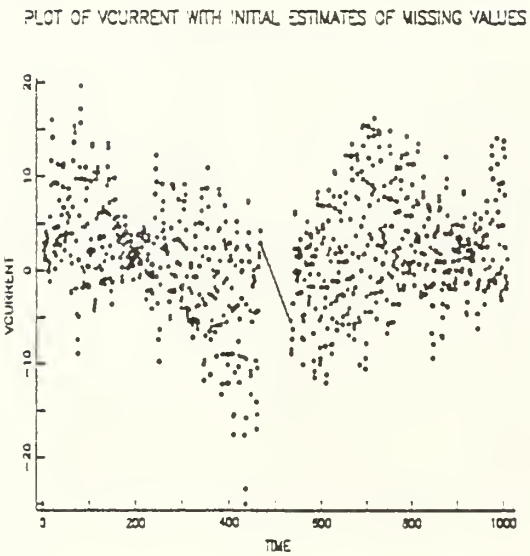

Figure 4: V-component of Current Velocity (missing values linearly interpolated)

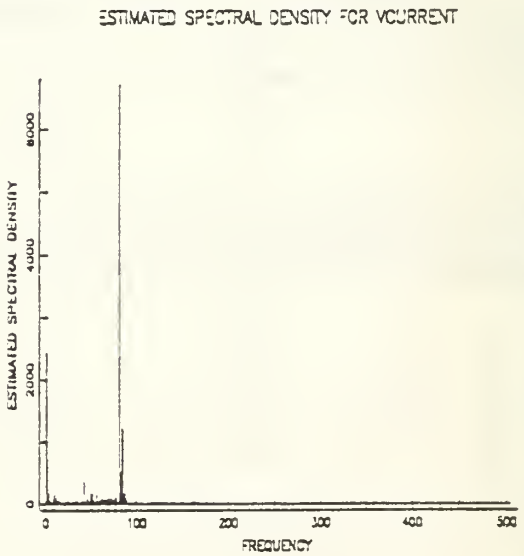

Figure 6: Periodogram of V-component of Current Velocity after linear detrending

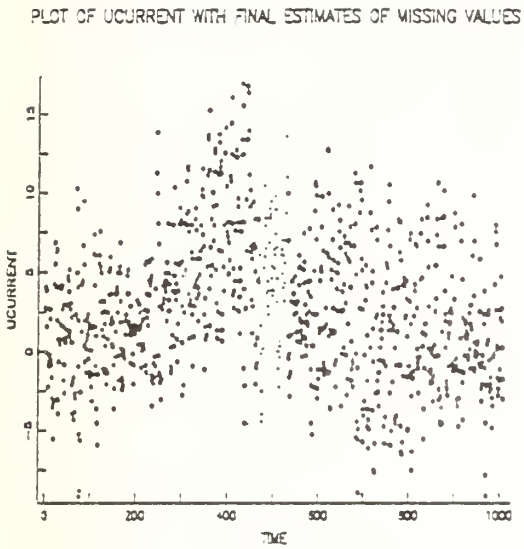

Figure 7: U-component of Current Velocity (missing values estimated using complete interpolation procedure)

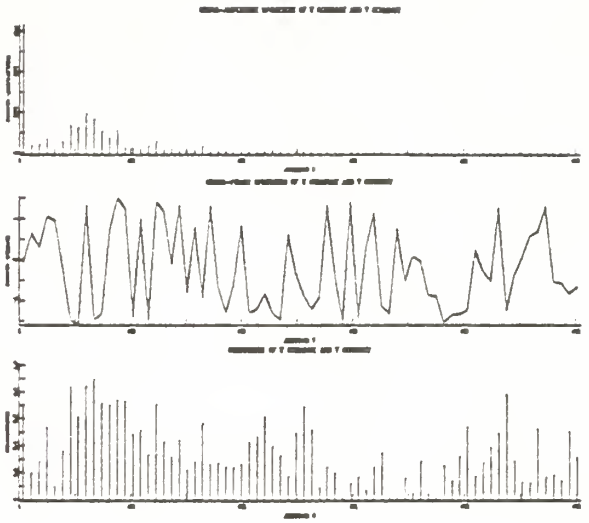

Figure 9: Cross-amplitude Spectrum (top), Phase Spectrum (middle), and Coherence (bottom) of U-component and V-component of Current Velocity

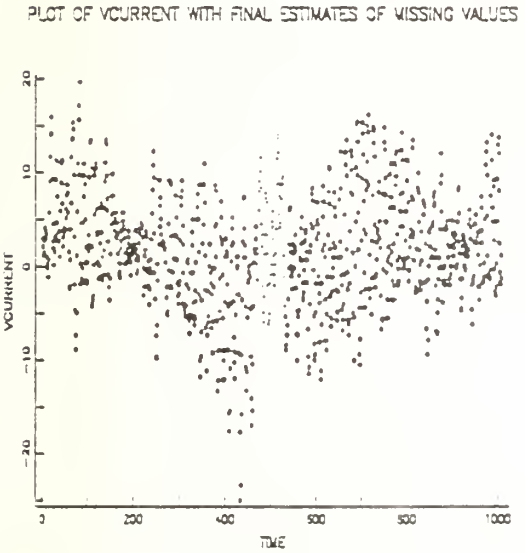

Figure 8: V-component of Current Velocity (missing values estimated using the complete interpolation procedure)

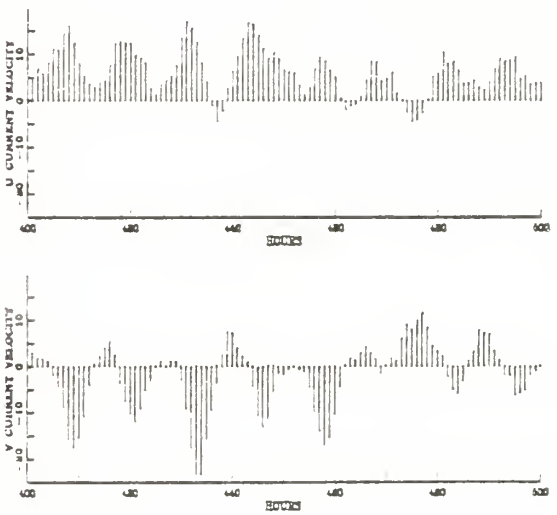

Figure 10: Detailed segment of U-component and V-component of Current Velocity after final estimation of missing values

INITIAL DISTRIBUTION LIST

1. Library (Code 0142) 2
   Naval Postgraduate School
   Monterey, CA 93943-5000

3. Defense Technical Information Center 2
   Cameron Station
   Alexandria, VA 22314

4. Office of Research Administration 1
   Code 012A
   Naval Postgraduate School
   Monterey, CA 93943-5000

5. Prof. Peter Purdue, Code OR/Pd 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

6. Prof. Peter A. W. Lewis 20
   Code OR/Lw
   Naval Postgraduate School
   Monterey, CA 93943-5000